

Be In The Know: Connecting News Articles to Relevant Twitter Conversations

Bichen Shi, Georgiana Ifrim, Neil Hurley

Insight Centre for Data Analytics
University College Dublin
Belfield, Dublin 4, Ireland

Abstract

In the era of data-driven journalism, data analytics can deliver tools to support journalists in connecting to new and developing news stories, e.g., as echoed in micro-blogs such as Twitter, the new citizen-driven media. In this paper, we propose a framework for tracking and automatically connecting news articles to Twitter conversations as captured by Twitter hashtags. For example, such a system could alert journalists about news that get a lot of Twitter reaction, so that they can investigate those conversations for new developments in the story, promote their article to a set of interested consumers, or discover general sentiment towards the story. Mapping articles to appropriate hashtags is nevertheless very challenging, due to different language styles used in articles versus tweets, the streaming aspect of news and tweets, as well as the user behavior when marking certain tweet-terms as hashtags. As a case-study, we continuously track the RSS feeds of Irish Times news articles and a focused Twitter stream over a two months period, and present a system that assigns hashtags to each article, based on its Twitter echo. We propose a machine learning approach for classifying and ranking article-hashtag pairs. Our empirical study shows that our system delivers high precision for this task.

Introduction

Since its start in 2006, Twitter has established itself as an alternative media source, with its 500 million users sending more than 500 million tweets daily on every possible topic. The 140 character messages called tweets, are typically grouped around the same subject by dedicated hashtags, e.g., political events: #btw13 (German Elections); crises: #Egypt, #USshutdown; natural disasters: #Haiyan (Philippines typhoon); epidemics: #H1N1; sports: #worldcup, #ashes; celebrities: #Messi, #royalbaby. Sometimes the news flows from the mainstream media to Twitter, and sometimes it is the Twitter users that first set the ground for breaking news stories, e.g., the 2009 airplane landing on the river Hudson.

Increasingly, Twitter conversations and calls to action that mobilize masses have dedicated hashtags, as showcased by recent world events, e.g., #ArabSpring, #Syria,

#freethe7. Twitter hashtags thus lead to the formation of ad hoc publics around specific themes and topics without the need for the users to be otherwise explicitly connected (Bruns and Burgess 2011). In our study we learned, for example, that for popular sporting events involving the national teams in Ireland, Twitter users prefer the hashtag #coybig (short for Come On You Boys in Green). For a recent political scandal in Germany, related to BMW funding Angela Merkel's political campaign, one of the hashtags preferred by users for commenting on this news was #buy_a_merkel. Hashtags can convey information about the community that uses them or the sentiment of the messages they group. In the US, for example, the hashtag #RacismEndedWhen groups tweets that mock a GOP tweet that seemed to suggest that racism had been ended. For an outsider, or even for an insider that doesn't continuously track the massive Twitter activity, it is close to impossible to stay in the know when it comes to the right hashtags or users to follow, for current and developing news stories. Nevertheless for journalists in particular, it is vital to get to the right hashtags quickly, in order to be able to follow new developments on topics of interest. Data analytics techniques can provide tools that link news stories to the relevant Twitter conversations.

Automatically mapping news articles to appropriate hashtags (where a hashtag is seen as a group of tweets forming a conversation around it) can be very challenging. This is due to different language styles used in the two types of data (e.g., clean, long articles versus messy, short tweets), the fast paced streaming aspect of both news and tweets (matching two streams moving at different speeds), as well as user behaviour when coining certain tweet-terms as hashtags. To showcase the third issue, in Table 1 we present an example news article and the categories we identified for the hashtags retrieved for it, in an initial pre-processing stage. The article is about Irish politics, in particular, the 2013 referendum to adopt a unicameral parliamentary system by abolishing one of the current two houses of parliament, the Seanad. This referendum was proposed by the Irish prime minister, Enda Kenny, and his party Fine Gael (FG). The hashtags retrieved for this article in an initial pre-processing step, range from highly specific and relevant (e.g., #seanad, #ireland-saysno), to general but still relevant (e.g., #ireland, #irishpolitics, #rtept for RTE Prime Time, a TV show broadcasting

a debate about the Seanad referendum), to abusive but potentially relevant (#caughtrotten referring to Irish PM Enda Kenny's role in this referendum), to irrelevant (e.g., #tobaccodirective, #mentalhealth, referring to other topics). We can see from this example that an approach that can accurately filter irrelevant hashtags and rank relevant hashtags can deliver value by connecting to the right Twitter conversations.

In this paper we propose a framework for connecting news articles from mainstream media to their echo on the Twitter stream. We discuss the data collection process for continuously gathering, processing and connecting a stream of news articles and a focused Twitter stream relevant to the tracked news stories. We analyze relevant features and propose a machine learning algorithm for ranking hashtags for a given news story. Our experiments show that our system can achieve high precision on this task. The rest of the paper is organized as follows. Section 2 discusses related work and our contributions. In Section 3 we explain the data collection process, while in Section 4 we describe the process of modeling hashtag ranking as a learning problem. In Section 5 we discuss our results and Section 6 concludes with directions for future work.

Related Work

Recent years have seen an explosion of research work analyzing social media (e.g., most prominently the micro-blog Twitter) and the connection between traditional media and this new form of reporting. Twitter studies focus on topics such as detecting political leaning from tweets (Boutet et al. 2012; Himelboim, McCreery, and Smith 2013), sentiment and opinion mining (Pak and Paroubek 2010; Liu and Zhang 2012; Luo, Osborne, and Wang 2012), summarising sporting, economic and other events using Twitter (Popescu and Pennacchiotti 2011; Nichols, Mahmud, and Drews 2012; Hu et al. 2012); analysing news spread through the network (Kunegis and Alhadi 2011; Artzi, Pantel, and Gamon 2012); content curation (Agichtein et al. 2008; Castillo, Mendoza, and Poblete 2011); user influence and authority (Bakshy et al. 2011; Romero et al. 2011); detecting breaking news stories from the massive tweet stream, potentially ahead of traditional media reporting (Sankaranarayanan et al. 2009; Kwak et al. 2010).

Among the diverse investigations of Twitter data, two categories are most relevant to this paper, namely, research that has focussed on hashtag recommendation or retrieval and studies considering the mapping between news articles and microblogs.

Hashtag Recommendation. Tag recommendation for tagging systems such as Last.fm and Delicious has been studied in a number of works such as (Krestel, Fankhauser, and Nejd 2009) that applies topic modelling using Latent Dirichlet Allocation (LDA) to the problem. Focusing in particular on hashtag retrieval over a Twitter corpus, in (Efron 2010), language modelling is used to find hashtags given a keyword query. A model

of each hashtag is learned from the set of tweets that contain the tag as a multinomial distribution over terms. Hashtags are ranked according to the KL divergence of their corresponding model to the query model. The related issue of keyphrase extraction from Twitter is studied in (Zhao et al. 2011a). Rather than associate a set of existing hashtags with individual tweets, the goal in this work is to recommend keyphrases generated from the entire vocabulary. The method follows three steps. In the first step, a topic model of the tweets is built, using a modified version of LDA suitable for short documents, proposed in (Zhao et al. 2011b), in which a tweet is generated from just a single topic along with a background noise 'topic' learned from the entire corpus. Next, a pagerank-like algorithm is run over each topic, to identify the most influential terms associated with the topic. Finally, keyphrases are generated by combining the top ranked terms and are ranked according to their relevance and interestingness. In (Ding, Zhang, and Huang 2012) the issue of recommending hashtags to untagged tweets is addressed. An LDA topic model is used to categorise tweets into topics and a translation probability maps topics to hashtags. The method is modified in (Ding et al. 2013) by replacing standard LDA with the topic model of (Zhao et al. 2011b).

News and Tweets. Work that investigates the connection between news and Twitter includes (Štajner et al. 2013). Given a set of tweets that specifically mention the URL of a given article, this work focuses on a method to filter this set into a subset of most interesting tweets. The authors use four indicators of interestingness, namely informativeness, opinionatedness, popularity and authority to filter the initial set. TweetMogaz (Magdy 2013), a system for microblog search and filtering, aims to find tweets relevant to regional news. It relies on a curated list of *key players* from which to collect an initial set of relevant tweets. The initial set is augmented, by firstly extracting a set of keywords from news sites and searching for tweets containing these keywords. The *keyword* tweets are filtered by training a classifier using the *key player* tweets as positive examples and a set of random tweets as negative examples. The *keyword* tweets that are classified as positive are retained. In (Lehmann et al. 2013), a system is proposed to support journalists in rapidly detecting follow-ups to their news articles that break on Twitter. Here again, the data extraction process starts with a set of tweets that mention the URL of a given news article. From this initial set of tweets, a set of users that tweeted the article within a limited time after its first tweet is gathered, and the tweets of this user group or *crowd* over a following time-period are analysed to discover new tweets related to the story. Other works investigated automatic news detection from tweets (Subašić and Berendt 2011), recommending news articles using tweets (Phelan et al. 2011), forecasting the popularity of news using Twitter (Bandari, Asur, and Huberman 2012), or enhancing news articles with information extracted from Twitter, such as *comment tweets* (Kothari et al. 2013).

Our work differs from the above research in a number of ways. In particular, we address hashtag recommendation in a streaming context, with a requirement that the model be up-

Table 1: Example news article and initially retrieved hashtags (before learning algorithm is applied).

News Article	Retrieved Hashtags (no learning)	Hashtag Category
Headline: <i>FG fears day of reckoning over Enda Kenny's Seanad gamble</i>	#seanad, #irelandsaysno, #enda	Relevant (Specific)
	#ireland, #rtept, #news	Relevant (General)
Sub-headline: <i>There is deep concern within the Fine Gael ranks that its populist referendum campaign misfired so badly</i>	#caughttrotten, #whip	Relevant (Abusive)
	#tobaccodirective, #mentalhealth	Irrelevant

dated on a daily basis. Rather than apply topic modeling on a large, static Twitter corpus, containing potentially many diverse topics, we attempt to filter irrelevant tweets directly by using the news articles to be hashtagged in order to focus the data collection from the Twitter stream. Nevertheless, unlike other work on connecting articles and microblogs, we avoid seeding our data collection with a curated user group or with tweets that specifically mention the articles in question (via the URL). As discussed later, our dynamic-keyword Twitter stream allows for a wide set of tweets to be gathered, while ensuring that the collection contains relevant tweets with high probability. We believe that our search strategy provides sufficient breadth to allow high recall in gathering relevant hashtags, while avoiding being drowned in a vast sea of Twitter noise. We alternate this high recall with a high precision oriented step, by using a learning approach to rank the retrieved hashtags for each article.

Our contributions are as follows: (1) we propose a focussed Twitter data collection strategy based on dynamic keyword extraction from news articles; (2) we formulate a learning algorithm for assigning hashtags to news articles; (3) we deliver a system for matching a daily news stream and a relevant Twitter conversation stream.

Data Collection

All the data collected for this study is available upon request for research purposes.

News Articles from RSS Feeds

We gathered the news articles streamed on the Irish Times RSS feeds between October 7, 2013 and November 30, 2013, by polling the RSS feeds every 5 minutes, yielding a total of 4,862 unique articles. The Irish Times is an Irish mainstream media outlet, that covers Irish news and high impact world news. There are three RSS feeds for general news and more specialised business and sport news. Each article has a headline, a one paragraph description that summarises the article (sub-headline), and the article body. Although there is additional meta-data associated with each article, in the form of manually assigned topics and some named entity annotations, these tend to be scarce and noisy, thus we currently do not use this meta-data in our approach.

We use Python for scrapping the urls from the RSS feed, downloading the html files and processing the text. Table 2 shows statistics on the number of news articles retrieved daily from the Irish Times RSS feed. The minimum number of articles corresponds to a Sunday (October 27, 2013),

Table 2: Statistics on the daily number of articles in the Irish Times RSS stream.

Min.	Median	Mean	Max	Stdev
87	168	159	213	32.47

while the maximum corresponds to the release of the Irish budget for 2014 (October 15, 2013). To obtain at-a-glance coverage of the newsfeed, needed for collecting a focused Twitter stream (as explained below), we extract representative keywords for each downloaded article. We parse the headline and sub-headline, part-of-speech tag this text, and extract nouns and named entities using shallow parsing techniques and heuristics (e.g., we extract Aer Lingus, Enda Kenny, etc.). We do not use the article-body for keyword extraction, since it poses risks of topic drift and noise. For example, for the news article in Table 1 we extract the keywords *enda kenny, fine gael, fg, fears, seanad*.

Focused Twitter Stream

We have experimented with several strategies for collecting Twitter data relevant to the daily news stream. Since we are interested in continuously streaming news and corresponding tweets, we use the Twitter Streaming API¹. The Streaming API can be employed with either keywords (words or phrases), geographical boundary boxes or user ID. Studies show (Morstatter et al. 2013) that the Twitter Streaming API provides access to 1% up to 40% of the public tweet-stream². This data may nevertheless be irrelevant to our focused information need. We looked at 4 alternatives to gather Twitter streaming data: using a curated set of users (200 Irish journalists), a static set of keywords (names of cities in Ireland), a geo-focused stream (using location coordinates to capture Irish tweets) and a dynamic set of keywords extracted from the stream of RSS news articles every 30 mins each day. Comparing the 4 streams, the dynamic-keyword stream was considerably larger, with a total of 23,362,818 unique tweets, as compared to 291,141 in the curated user stream, 1,629,678 in the geo-stream, and 8,527,952 in the static keyword stream. In this work we focus on the fourth method: the dynamic-keyword-focused Twitter stream. Due to the restriction of the Streaming API to use a maximum of 400 keywords, we limit the number of keywords extracted

¹<https://dev.twitter.com/docs/streaming-apis>

²<http://www.brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care>

Table 3: Processed keyword set by permutation.

Original keywords/phrases	Final keywords/phrases
enda kenny	enda kenny
fine gael	fine gael
fg	fg fears
fears	fg seanad
seanad	fears seanad

Table 4: Statistics on the daily number of tweets in our focused (dynamic-keyword) Twitter stream.

Min.	Median	Mean	Max	Stdev
91,828	350,833	424,778	1,389,265	245,536.3

from articles by giving preference to named entities and frequent nouns.

Additionally, we noticed that in order to get relevant tweets, it helps if we constrain each tweet returned by the Twitter API to contain at least two article keywords. We achieve this by splitting our original keyword set, into individual keywords, and creating all possible permutation pairs as our final keyword set, with the constraint that we freeze named entities. For example, for the article in Table 1, we process the keyword set *enda kenny*, *fine gael*, *fg*, *fears*, *seanad* by keeping the named entities and permuting the single keywords to form pairs, as shown in Table 3. We apply this process every 30 minutes to *all* the RSS articles downloaded up to that point in time, pool the keywords together, and re-connect with the Streaming API using the updated keyword list.

Through this process we aim to retrieve a large set of relevant tweets whilst not being restricted to a set of manually curated user lists, locations or keywords. The problem of retrieving relevant tweets to a set of news has been pointed out in recent research (Kothari et al. 2013) with ad-hoc retrieval techniques achieving low Recall (0.5). Prior work relies mostly on tweets where the url of the article is explicitly provided, therefore obtaining a clean but potentially small set of tweets. Our initial tweet-retrieval process gathers a large set of potentially relevant tweets, which we carefully filter in a following step, using a machine learning approach. Since we aim at capturing and continuously tracking active conversations around particular news stories, it is important to not have to artificially restrict the set of tweets from which we extract hashtags. Table 4 shows statistics on the number of tweets in the daily Twitter streams over the 2 months data. The minimum tweet activity happened on Sunday, October 20, 2013, and the maximum activity corresponds to October 30, 2013, an eventful day for Champions League (european soccer competition). Table 5 gives a sample of tweets from the tweet-bag of our example article.

Learning Algorithm for Scoring Hashtags

In this section we discuss the process of modeling hashtag selection as a learning problem. We parse the stream of news articles and the Twitter stream daily, in order to extract and rank hashtags for each news article. For tweet-

Table 5: Example tweets with hashtag #seanad, from the tweet-bag of Table 1 article.

Like many of his cabinet, the receive knob is broken, so he's permanently on transmit #enda #seanad

Poor Enda Kenny @MayoGAA #stillhurting "We know now that like the All Ireland Final, it is not going to be replayed... #seanad

Jaysus Inda Kenny says no #seanadref rerun. So, only feckin thing we've 2 look forward 2 is Fine Gael night of d'long knives #vinb #seanad

So Kenny decides today to reform the Seanad after spending millions on a referendum-what an ejit we have running this country#waste#seanad

FG fears day of reckoning over Enda Kenny #Seanad gamble - The Irish Times - Mon, Oct 07, <http://t.co/8FtnVxV86d>

processing, we remove stop words, punctuation, URLs and user names, and apply stemming to the remaining terms. For each day, and each news article, we separate the tweets of the corresponding Twitter stream *per article*, based on a shallow matching of tweet keywords and article-keywords (as extracted for the Streaming API and showcased in Table 3). This results in a local tweet-bag per article, that can be analyzed for extracting hashtags and hashtag information, e.g., frequency, keyword-profile describing the hashtag as reflected in its tweet-bag. Next, we form article-hashtag pairs, and compute features of each pair useful for discriminating whether a hashtag is relevant to a given article.

Features. Currently, we extract four features for each article-hashtag pair, two features that characterize the local hashtag profile, while the other two characterize the global hashtag profile, useful for describing specific versus general hashtags, as shown in Table 6.

One of the first features we select is the cosine similarity between the tf.idf keyword profile of the article, and that of the local hashtag profile (as extracted from the tweet-bag). To avoid noise in the article profile, rather than selecting terms from the full article-body, we only select them from the headline and sub-headline, but compute their tf.idf weight using the entire article. Additionally, we extract the local popularity of the hashtag, i.e., the number of tweets in the article tweet-bag, mentioning that hashtag.

We extract the global frequency of the hashtag in the entire Twitter stream (rather than only the local tweet-bag of the article), and we compute the cosine between the local and the global hashtag keyword-profile, to assess how much does the global hashtag profile diverge from the local profile. Note that globally, the same hashtag may refer to different events, or a hashtag may be preferred over a time window to refer to a certain event, and then slowly discarded or outweighed by other hashtags. Therefore, using local and global features for each hashtag, addresses the issue of time-of-use

Table 6: Features of learning algorithm for scoring hashtags.

Features	Definition
Local Frequency	The number of tweets in the article tweet-bag, mentioning that hashtag.
Local Cosine Similarity	Cosine similarity between the tf.idf keyword profile of the article, and that of the local hashtag profile.
Global Frequency	The number of tweets in the entire Twitter stream, mentioning that hashtag.
Global Cosine Similarity	Cosine similarity between the local and the global tf.idf hashtag keyword profile (measures how specialized a hashtag is).

Table 7: Statistics on articles and article-hashtag pairs extracted for the manually labeled dataset.

Date	Total articles	With hashtags	Total pairs
Oct23	146	90	1,107
Nov23	142	87	1,395

and scope of a hashtag. For each article-hashtag pair, we now have four features describing how relevant a hashtag may be for a given article. We normalize all four features to the $[0, 1]$ interval. Next, we discuss how to use these features and a set of manually labeled article-hashtag pairs for learning to identify relevant hashtags.

Labeled Data. In order to build a classification algorithm for recognizing relevant hashtags, we need labeled article-hashtag pairs. We selected two days at random from the two month dataset, October 23, and November 23, 2013, extracted all the article hashtag pairs and their features as described above, and asked two annotators to manually label each pair. Table 7 shows statistics on the number of article-hashtag pairs extracted. The data annotation process is very intensive, since for many pairs it is hard to assess whether the hashtag is relevant or not, without reading the article fully, and searching Twitter for the particular hashtag, or closely analyzing the hashtag-local-profile. In our study, we have found several categories of hashtags, as shown in Table 1. Given the difficulty of this task, we asked the annotators to decide which of the three scenarios applies to each pair: (1) a hashtag is *specific and relevant* to the topic of the news article (e.g., #seanad, #seanref, for the example article in Table 1), (2) *general and relevant* (e.g., #rtepm, #news), or (3) *irrelevant* (e.g., #tobaccodirective, is related to Enda Kenny, but on a different topic). For abusive hashtags, the annotators were advised to decide depending on the local context, e.g., #caughttrotten is relevant to our example article since it groups users discussing the impact of the negative referendum result on the leading party Fine Gael. For the purpose of our experiments, we merged the first two classes into simply relevant (a positive example in binary classification) or irrelevant (negative example). The inter-annotator agreement was 80%. We used the subset of examples where both annotators agreed for training/testing a classification algorithm.

Classification Algorithm We train and test our approach by employing a series of Weka (Witten, Frank, and Hall 2011) classification algorithms. The algorithm only sees the examples as described by the four features, and can learn thresholding strategies on the provided features. For example, to

classify a hashtag as relevant for a given article, a classification algorithm may learn (from the training set) that the cosine feature should be higher than 0.5 and the hashtag frequency should be close to 1. Additionally, most classifiers provide a score describing the likelihood that a hashtag is relevant to the article. We use this classification score to rank hashtags for each article.

Evaluation

In order to evaluate our overall strategy for retrieving, learning, and ranking hashtags, we present three evaluation settings.

- **Small:** We use the manual labeling of article-hashtag pairs for two random days October 23, 2013 (90 articles with hashtags) for training, and November 23, 2013 (87 articles with hashtags), for testing. There are 874 training examples (39.7% positives, 61.3% negatives) and 1,122 test examples (45.7% positives, 54.3% negatives).
- **Medium:** Many of the Irish Times articles are discussed on Twitter, by posting the URL (or short URL) of the article directly in the tweet. We parse all tweets from our focused Twitter stream, containing URLs of Irish Times articles (we expand the short URL and only retain genuine URLs pointing to real Irish Times articles), and extract the Twitter-user-assigned hashtags for those articles. We then use the user-assigned-hashtag data as a form of ground truth, by assuming all user-assigned hashtags are relevant. We train a classifier on the manually labeled data, and check how many user-hashtags can our learning algorithm recognize. There are 1,136 articles with user-hashtags, that lead to 2,773 article-user-hashtag pairs, extracted from 2,801 tweets. In the intersection between the article-user-hashtag pairs and RSS article-hashtag pairs, there are 732 articles, and 1,146 articles-user-hashtag test examples. Note that although this experiment covers many more articles than the previous one (732 versus 87), the user-hashtag set is smaller than the hashtag set retrieved by our technique, therefore the test data only contains 1,146 pairs. We train on 2,502 example pairs (October 23, and November 23, 2013) and test on the 1,146 article-user-hashtag examples.
- **Large:** For each article in our two month set, we attempt to assign hashtags with the procedure described above. We train a classifier based on the manually labeled data and the user assigned hashtags (3,648 training examples), and test it on all the article-hashtag pairs retrieved, to assign a classification score to each article-hashtag pair.

There are 3,388 articles that get assigned hashtags and a total of 62,764 article-hashtag example pairs. For each article that gets relevant hashtags (i.e., it gets a classification score above 0.5), we rank the hashtags assigned, and take top-3 most relevant hashtags with respect to the classification score. This process results in 2,838 articles with at least one hashtag classified as relevant. Out of this set, we randomly select 422 articles for manual evaluation (about 15% of all articles with relevant hashtags). This results in 1,029 article-hashtag test pairs to be manually evaluated.

Error Metrics

We employ metrics from both machine learning and information retrieval to assess the quality of our results.

Classification. We compute the following standard binary classification metrics (Witten, Frank, and Hall 2011), where TP stands for true positive, FP for false positive, TN true negative and FN false negative:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + FP + TN + FN} \\ \text{Precision} &= \frac{TP}{TP + FP}; \text{Recall} = \frac{TP}{TP + FN} \\ F1 &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned}$$

The above metrics assume a fixed classification threshold, but one can typically vary this threshold to improve classifier performance (e.g., by tuning on the training set). Rather than focusing on a binary split into positives and negatives, other metrics characterise the ranking performance of a classifier. The ROC curve characterises the performance of a binary classifier, by capturing the fraction of true positives versus that of false positives at varying classification thresholds (Fawcett 2004). The area under the ROC curve (AUC) is an aggregate measure corresponding to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. An AUC of 1 represents perfect ranking of all positives above all negatives.

Information Retrieval. The above metrics assess the classification algorithm over all article-hashtag pairs. In order to assess the per article hashtag ranking quality, we also employ metrics from information retrieval. We evaluate the classifier-induced ranking of hashtags, for each article, by the Precision@1 and the Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen 2000; Wang et al. 2013) as defined below.

$$DCG@k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

$$NDCG@k = \frac{DCG@k}{IdealDCG@k}$$

These are standard information retrieval metrics for evaluating the quality of a ranking function (Manning, Raghavan, and Schütze 2008). The Precision@1 captures how satisfied the user is with the best ranked hashtag for each article. It is computed as the number of times that a relevant hashtag is in the first position of the ranking, weighted by the relevance score and

Table 8: **Small** (baselines): Top-3 relevant hashtags using most frequent or highest cosine.

Baselines	Precision
Most Frequent Top-3	0.548
Highest Cosine Top-3	0.634

normalized. The NDCG describes the cumulative gain the user obtains by examining the retrieval results up to a given rank position k . NDCG makes it possible to evaluate the ranking of hashtags uniformly across articles (independent of whether the article gets k or less relevant hashtags). It further allows a more fine-grained evaluation, by penalising more the relevant results placed at low ranks.

Results: Small Experiment

For the first experimental setting, we had 874 training examples (October 23, 2013) and 1,122 test examples (November 23, 2013) on which the labels of both annotators agreed.

Baselines. In order to assess the actual utility of a learning approach, we first evaluate two simple baseline techniques. On the test set (November 23, 2013), we select the top-3 hashtags per article (257 pairs out of 1,122), using the highest local hashtag frequency and the highest local cosine similarity. Table 8 shows the precision of results using these two simple heuristics.

Learning Approach. We now evaluate the classifier’s ability to retrieve all the hashtags deemed relevant by our annotators as well as its ability to rank them before the irrelevant ones. We experimented with a series of Weka classifiers, with default parameter settings. MultilayerPerceptron, Logistic (regularised logistic regression) and Kstar (K-nearest neighbours with entropy-based-distance) delivered the best results, as shown in Table 9. We note that all three classifiers have high precision (0.85), recall (0.80) and AUC (0.92), showing that the classifier ranks relevant hashtags before irrelevant ones. The AUC is particularly important, since ultimately it is useful to rank the hashtags of each document, from most relevant to least relevant. We also experimented with swapping training and test (i.e., training on November 23, and testing on October 23), and the results are very similar. We note that the Logistic classifier had the highest Precision and AUC. Additionally, the Logistic classifier is a linear model that can be easily interpreted and its classification scores are true probabilities. The Logistic model deemed all four features as important (non-zero weights), with the local cosine feature getting the highest weight, followed by the frequency based features, and ending with the global cosine.

Results: Medium Experiment

In this setting, we train classifiers on the manually labeled data from the **Small** setting, and test them on Twitter-user-hashtagged data. As training data we analyze three settings, October 23, 2013, November 23, or both days together as training, and article-user-hashtag pairs as test (1,146 test examples). Note that we assume that all the user-assigned hashtags are relevant, which may not necessarily be the case, since sometimes users also assign spurious hashtags, e.g.,

Table 9: **Small** (learning): Hashtag classification on manually labeled test set.

Weka Classifier	Accuracy	Precision	Recall	F1	AUC
MultilayerPerceptron	84.6%	0.850	0.807	0.846	0.921
Logistic	84.4%	0.876	0.770	0.844	0.924
Kstar	83.9%	0.861	0.774	0.839	0.911

Table 10: **Medium** (baselines): Top-3 relevant hashtags using most frequent or highest cosine.

Baselines	Recall
Most Frequent Top-3	0.503
Highest Cosine Top-3	0.644

Table 11: **Medium** (learning): Hashtags identified as relevant on the Twitter-user-labeled dataset.

Training Data	Weka Classifier	Recall
Oct23	MultilayerPerceptron	0.781
	Logistic	0.756
	Kstar	0.704
Nov23	MultilayerPerceptron	0.792
	Logistic	0.808
	Kstar	0.787
Oct23 & Nov23	MultilayerPerceptron	0.845
	Logistic	0.797
	Kstar	0.776

#annoying #omg. Our algorithm may consider such hashtags as irrelevant, in particular if the global hashtag profile strongly diverges from the local hashtag profile. Since we assume all the test examples are relevant (there are no negative test examples), in this setting the Accuracy is the same as Recall. We again test the two baselines to check Recall at top-3, as shown in Table 10. In Table 11 we show Recall over all Twitter-user-hashtags retrieved by our algorithm as being relevant (classification score above 0.5). We note that when training only on October or November 23, the classifier retrieves about 80% of user-hashtags as relevant, while when we increase the amount of training data, by combining the October and November examples, the accuracy of MultilayerPerceptron stands at 84.5%, a similar value to that of the small setting experiment.

Results: Large Experiment

Building on observations from the previous two evaluation settings, we train and apply a classifier to the entire two month data collection. For training we use the labeled examples of October 23, November 23, and the Twitter-user-hashtagged examples. We then apply this classifier to all the article-hashtag pairs extracted from the RSS and Twitter streams, a total of 62,764 test examples (3,388 unique articles). We extract only articles that get at least one relevant hashtag (based on classification score above 0.5; 2838 articles), and manually assess a random sample (422 articles, 1,029 pairs), using 0 and 1 relevancy scores. We evaluate both the filtering quality, i.e., the classification across all

article-hashtag pairs (to assess the Precision over the pairs classified as relevant), as well as the hashtag ranking quality per article, using information retrieval metrics. For the article oriented metrics, we use Precision@1 and NDCG@3 and average them across all articles. We show the average values and the t-test 95% confidence interval for Precision@1 and NDCG@3 across all articles in Table 12. We note that the precision for the filtering step (binary classification into relevant/irrelevant) is fairly high (Precision 0.86), and similar to what we have seen in the previous experiments. When we evaluate the quality of ranking of hashtags for each article, we see a similar result: the Precision@1 is 0.9, while the NDCG@3 which penalizes relevant hashtags ranked at low ranks, is 0.87.

Discussion

In order to make the whole methodology more explicit, in Table 13 we show some example articles from our annotated sample of the **Large** setting, their extracted keywords, their (up to top-3) ranked hashtags, together with the features extracted for the corresponding pair, the classifier score and the annotator relevance score. We observe that for local as well as international news (first 3 articles), the hashtags assigned and ranked (by classifier score) are relevant and quite specific (e.g., #ecb, #walshwhiskeydistillery).

We found three main reasons why an article does not get any (relevant) hashtag: the article-keyword extraction process is faulty (due to part-of-speech tagging errors or due to the fact that the extracted keywords are too generic); there is no discussion on Twitter about that particular news story; the tweets relevant to an article do not contain any hashtags. The aspect of assigning noisy or irrelevant hashtags can be mitigated to some extent by tuning the classifier threshold (here we used the default classification score of 0.5). Additionally, the four features describing each article-hashtag pair could be enhanced, e.g., using user authority to re-weight tweets, filtering spammy hashtags (e.g., #ff, #followback). Regarding the lack of hashtags in the tweet-bag of an article, in such cases we could employ recent techniques for extracting informative tweets (Štajner et al. 2013), or adapt our approach for the problem of assigning Twitter users (rather than hashtags) relevant to a given news article. Among the categories of news articles from our dataset that get a lot of relevant hashtags, are those related to sporting events. This is likely due to the fact that sporting events get a lot of reaction on Twitter. The manual annotation for the learning approach is potentially noisy, since at times it is quite difficult to decide whether a hashtag is relevant or not, without considerable background knowledge. In this respect we plan to employ crowd sourcing platforms such as Crowdfunder, in order to obtain larger and possibly cleaner labeled datasets.

Table 12: **Large** (learning): Hashtag filtering and ranking on two month article-dataset.

Precision (over 1,029 pairs)	Precision@1 (over 422 articles)	NDCG@3 (over 422 articles)
0.869	0.900 ([0.871, 0.929], $p < 2.2e - 16$)	0.877 ([0.850, 0.904], $p < 2.2e - 16$)

Table 13: Example results from our two-month annotated sample.

Article URL	Article Keywords	Hashtag	LFr	LCo	GFr	GCo	ClassifScor	RelScor
tech-titans-in-town-for-dublin-web-summit-1.1573638	dublin, dubstarts, summit, tech, web	#websummit	1.00	0.35	0.58	0.82	0.92	2
		#tech	0.23	0.45	0.70	0.37	0.89	1
		#web	0.42	0.41	0.40	0.52	0.82	2
whiskey-distillery-to-create-55-jobs-for-co-carlow-1.1562843	carlow, co, distillery, walsh, whiskey	#whiskey	1.00	0.73	0.16	0.56	0.99	2
		#carlow	0.90	0.64	0.16	0.53	0.99	2
		#walshwhiskeydistillery	0.66	0.61	0.00	1.00	0.97	2
ecb-s-draghi-moves-to-ease-fears-on-interest-rates-1.1604077	banks, draghi,ecb	#ecb	1.00	0.50	0.39	0.42	0.97	2
		#draghi	0.54	0.58	0.19	0.69	0.96	2
		#news	0.00	0.46	0.89	0.29	0.90	1
climate-change-watchdog-must-be-robust-and-independent-says-report-1.1601809	advisory, climate, council, expert, fiscal	#delleir	1.00	0.27	0.00	1.00	0.60	0
		#job	0.00	0.30	0.80	0.35	0.53	0
		#delcfe	0.89	0.26	0.00	1.00	0.51	0
europe-bank-payouts-capped-as-capital-bar-keeps-rising-1.1603937	capital-europe	#europe	0.79	0.35	0.55	0.02	0.84	1
		#travel	0.66	0.27	0.68	0.38	0.72	0

Table 14: Six example articles clustered by hashtag #seanad based on learning algorithm score.

Headline	Classif Score
Kenny embroiled in tense spat over referendum at FG meeting.	0.99
The political reform shortfall remains.	0.97
Taoiseach anxious to see Seanad become effective watchdog.	0.92
FG fears day of reckoning over Enda Kenny's Seanad gamble.	0.84
Government says Seanad reform is on the agenda.	0.83
Final arguments made in appeal against Calley decision.	0.82

As a toy example show-casing the potential applications of connecting articles and hashtags, in Table 14 we cluster news articles in hashtag space. In this particular example, we simply sort news articles that get the hashtag #seanad, by classifier score. We note that news articles belonging to the same news topic (Seanad referendum) group together, although some do not have any terms in common in the original article headlines.

Conclusion

In this work we present a framework for connecting news articles to their relevant Twitter conversations, as semantically grouped by Twitter hashtags. We discuss the aspect of continuously tracking a stream of news and tweets, and present an approach for obtaining a large focused Twitter stream automatically seeded by a dynamic keyword set extracted from the articles. Furthermore, we model the problem of hashtag assignment as a classification problem, and analyze a frame-

work for hashtag retrieval and appropriate features and data for building a hashtag classifier. We evaluate our methods and show that our approach achieves high precision for this task.

Future Work. We plan to extend our study to track several RSS news feeds and Twitter conversations, and test a prototype with journalists. We also intend to investigate applications of our methods to clustering of articles in hashtag space, story tracking and event detection.

References

- [Agichtein et al. 2008] Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *WSDM*.
- [Artzi, Pantel, and Gamon 2012] Artzi, Y.; Pantel, P.; and Gamon, M. 2012. Predicting responses to microblog posts. In *NAACL HLT*.
- [Bakshy et al. 2011] Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *WSDM*.
- [Bandari, Asur, and Huberman 2012] Bandari, R.; Asur, S.; and Huberman, B. A. 2012. The pulse of news in social media: Forecasting popularity. In *ICWSM*.
- [Boutet et al. 2012] Boutet, A.; Kim, H.; Yoneki, E.; et al. 2012. What's in your tweets? i know who you supported in the uk 2010 general election. In *ICWSM*.
- [Bruns and Burgess 2011] Bruns, A., and Burgess, J. E. 2011. The use of twitter hashtags in the formation of ad hoc publics. European Consortium for Political Research Conference.
- [Castillo, Mendoza, and Poblete 2011] Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *WWW*.

- [Chakrabarti and Punera 2011] Chakrabarti, D., and Punera, K. 2011. Event summarization using tweets. In *ICWSM*.
- [Ding et al. 2013] Ding, Z.; Qiu, X.; Zhang, Q.; and Huang, X. 2013. Learning topical translation model for microblog hashtag suggestion. In *IJCAI*.
- [Ding, Zhang, and Huang 2012] Ding, Z.; Zhang, Q.; and Huang, X. 2012. Automatic hashtag recommendation for microblogs using topic-specific translation model. *COLING*.
- [Efron 2010] Efron, M. 2010. Hashtag retrieval in a microblogging environment. In *SIGIR*.
- [Fawcett 2004] Fawcett, T. 2004. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*.
- [Himelboim, McCreery, and Smith 2013] Himelboim, I.; McCreery, S.; and Smith, M. 2013. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication*.
- [Hu et al. 2012] Hu, Y.; John, A.; Seligmann, D. D.; and Wang, F. 2012. What were the tweets about? topical associations between public events and twitter feeds. In *ICWSM*.
- [Järvelin and Kekäläinen 2000] Järvelin, K., and Kekäläinen, J. 2000. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR*.
- [Kothari et al. 2013] Kothari, A.; Magdy, W.; Kareem Darwish, A. M.; and Taei, A. 2013. Detecting comments on news articles in microblogs. *ICWSM*.
- [Krestel, Fankhauser, and Nejdl 2009] Krestel, R.; Fankhauser, P.; and Nejdl, W. 2009. Latent dirichlet allocation for tag recommendation. In *RecSys*.
- [Kunegis and Alhadi 2011] Kunegis, N. N. T. G. J., and Alhadi, A. C. 2011. Bad news travel fast: A content-based analysis of interestingness on twitter. *WebSci*.
- [Kwak et al. 2010] Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *WWW*.
- [Lehmann et al. 2013] Lehmann, J.; Castillo, C.; Lalmas, M.; and Zuckerman, E. 2013. Transient news crowds in social media. In *ICWSM*.
- [Liu and Zhang 2012] Liu, B., and Zhang, L. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*.
- [Long et al. 2011] Long, R.; Wang, H.; Chen, Y.; Jin, O.; and Yu, Y. 2011. Towards effective event detection, tracking and summarization on microblog data. In *Web-Age Information Management*.
- [Luo, Osborne, and Wang 2012] Luo, Z.; Osborne, M.; and Wang, T. 2012. Opinion retrieval in twitter. In *ICWSM*.
- [Magdy 2013] Magdy, W. 2013. Tweetmogaz: A news portal of tweets. In *ACM SIGIR*.
- [Manning, Raghavan, and Schütze 2008] Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to information retrieval*.
- [Marcus et al. 2011] Marcus, A.; Bernstein, M. S.; Badar, O.; Karger, D. R.; Madden, S.; and Miller, R. C. 2011. Twitter: aggregating and visualizing microblogs for event exploration. In *SIGCHI*.
- [Morstatter et al. 2013] Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *ICWSM*.
- [Nichols, Mahmud, and Drews 2012] Nichols, J.; Mahmud, J.; and Drews, C. 2012. Summarizing sporting events using twitter. In *ACM IUI*.
- [Pak and Paroubek 2010] Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*.
- [Phelan et al. 2011] Phelan, O.; McCarthy, K.; Bennett, M.; and Smyth, B. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *Advances in Information Retrieval*.
- [Popescu and Pennacchiotti 2011] Popescu, A.-M., and Pennacchiotti, M. 2011. "dancing with the stars," nba games, politics: An exploration of twitter users' response to events. In *ICWSM*.
- [Romero et al. 2011] Romero, D. M.; Galuba, W.; Asur, S.; and Huberman, B. A. 2011. Influence and passivity in social media. In *Machine learning and knowledge discovery in databases*.
- [Sankaranarayanan et al. 2009] Sankaranarayanan, J.; Samet, H.; Teitler, B. E.; Lieberman, M. D.; and Sperling, J. 2009. Twitterstand: news in tweets. In *SIGSPATIAL*.
- [Štajner et al. 2013] Štajner, T.; Thomee, B.; Popescu, A.-M.; Pennacchiotti, M.; and Jaimes, A. 2013. Automatic selection of social media responses to news. In *ACM SIGKDD*.
- [Subašić and Berendt 2011] Subašić, I., and Berendt, B. 2011. Peddling or creating? investigating the role of twitter in news reporting. In *Advances in Information Retrieval*.
- [Wang et al. 2013] Wang, Y.; Wang, L.; Li, Y.; He, D.; and Liu, T.-Y. 2013. A theoretical analysis of ndcg type ranking measures. In *JMLR Proceedings*.
- [Whiting et al. 2012] Whiting, S.; Zhou, K.; Jose, J.; Alonso, O.; and Leelanupab, T. 2012. Crowdtiles: presenting crowd-based information for event-driven information needs. In *CIKM*.
- [Witten, Frank, and Hall 2011] Witten, I. H.; Frank, E.; and Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*.
- [Zhao et al. 2011a] Zhao, W. X.; Jiang, J.; He, J.; Song, Y.; Achananuparp, P.; Lim, E.-P.; and Li, X. 2011a. Topical keyphrase extraction from twitter. In *ACL HLT*.
- [Zhao et al. 2011b] Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; and Li, X. 2011b. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*.